A Geometric Perspective on Minimal Peer Prediction†

RAFAEL FRONGILLO, CU Boulder
JENS WITKOWSKI, ETH Zurich

Minimal peer prediction mechanisms truthfully elicit private information (e.g., opinions or experiences) from rational agents without the requirement that ground truth is eventually revealed. In this paper, we use a geometric perspective to prove that minimal peer prediction mechanisms are equivalent to power diagrams, a type of weighted Voronoi diagram. Using this characterization and results from computational geometry, we show that many of the mechanisms in the literature are unique up to affine transformations. We also show that classical peer prediction is "complete" in that every minimal mechanism can be written as a classical peer prediction mechanism for some scoring rule. Finally, we use our geometric characterization to develop a general method for constructing new truthful mechanisms, and we show how to optimize for the mechanisms' effort incentives and robustness.

Categories and Subject Descriptors: J.4 [Computer Applications]: Social and Behavioral Sciences—Economics; K.4 [Computers and Society]: Electronic Commerce

Additional Key Words and Phrases: Peer Prediction, Information Elicitation, Mechanism Design

1. INTRODUCTION

User-generated content is essential to the effective functioning of many social computing and e-commerce platforms. A prominent example is eliciting information through crowdsourcing platforms, such as Amazon Mechanical Turk, where workers are paid small rewards to do so-called human computation tasks, which are easy for humans to solve but difficult for computers. For example, humans easily recognize celebrities in images, whereas even state-of-the-art computer vision algorithms perform significantly worse.

While statistical techniques can adjust for biases or identify noisy users, they are appropriate only in settings with repeated participation by the same user, and when user inputs are informative in the first place. But what if providing accurate information is costly for users, or if users have incentives to lie? Consider an image annotation task (e.g. for search engine indexing), where workers may wish to save effort by annotating with random words, or words that are too generic (e.g. “animal”). Or consider a public health program that requires participants to report whether they have ever used illegal drugs, and where participants may lie about their drug use due to shame or eligibility concerns.

Peer prediction mechanisms address these incentive problems. They are designed to elicit truthful private information from self-interested participants, such as answers to the question “Have you ever used illegal drugs?” Crucially, peer prediction mechanisms cannot use ground truth. In the public health example this means the program cannot verify whether a participant has or has not ever used illegal drugs; it can only use the participants' voluntary reports.

The classical peer prediction method [Miller et al. 2005] addresses this challenge by comparing the reported information of a participant with that of another participant, and computing a payment rule which ensures that truth revelation is a strategic equilibrium. The major shortcoming of the classical peer prediction method with regard

†This paper is significantly extended from the paper “A Geometric Method to Construct Minimal Peer Prediction Mechanisms,” which appeared in the Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI’16).

Author’s addresses: raf@colorado.edu (Rafael Frongillo) and jensw@inf.ethz.ch (Jens Witkowski).
to practical applications is that it requires too much common knowledge. Bayesian Truth Serum mechanisms [Prelec 2004; Witkowski and Parkes 2012a; Witkowski 2014; Radanovic and Faltings 2013] relax these common knowledge assumptions but require participants to report a probability distribution in addition to the actual information that is to be elicited. That is, they are not minimal. The 1/p mechanism [Jurca and Faltings 2008; Jurca and Faltings 2011] and the Shadowing Method [Witkowski and Parkes 2012a; Witkowski 2014] relax the common knowledge assumptions of classical peer prediction to some degree while still being minimal.

Our Results. In this paper, we provide a complete characterization of the mechanism design space of minimal peer prediction, which includes the classical peer prediction method, output agreement, the 1/p mechanism, and the Shadowing Method as special cases. While it was known that every minimal mechanism requires some constraint on the agents’ belief models [Jurca and Faltings 2011], it was unknown which constraints allow for truthful mechanisms and how the constraints of different truthful mechanisms relate to one another. We answer these questions in Section 3 by adapting techniques from property elicitation [Lambert and Shoham 2009; Frongillo and Kash 2014], allowing us to prove the equivalence of minimal peer prediction mechanisms and power diagrams, a type of weighted Voronoi diagram. In Section 4, we then use this and results from computational geometry to show that all aforementioned mechanisms are unique with respect to their belief model constraints up to positive-affine transformation. One important corollary of this is that maximizing effort incentives for any of these mechanisms reduces to computing the effort-optimal positive-affine transformation. In Section 5, we show how to construct new truthful mechanisms for new conditions, followed by a detailed example of this construction in Section 6. In Section 7, we revisit the classical peer prediction method and show how to compute a mechanism that is maximally-robust with respect to deviations between the mechanism’s and the agents’ belief models. We conclude with useful directions for future work in Section 8.

2. PRELIMINARIES

In this section, we introduce the model, and review concepts in peer prediction and computational geometry.

2.1. Model

There is a group of \( n \geq 2 \) rational, risk-neutral and self-interested agents. When interacting with the environment, each agent \( i \) observes a signal \( S_i \), which is a random variable with values \( [m] := \{1, \ldots, m\} \) and \( m \geq 2 \). The signal represents an agent’s experience or opinion. The objective in peer prediction is to elicit an agent’s signal in an incentive compatible way, i.e. compute payments such that agents maximize their expected payment by reporting their signal to the mechanism (center) truthfully. To achieve this, all peer prediction mechanisms require that agent \( i \)’s signal observation tells her something about the signal observed by another peer agent \( j \neq i \). For example, this could be agent \( j = i + 1 \mod n \), so that the agents form a “ring,” where every agent is scored using the “following” agent. (Our results hold for any choice of peer agent.) Let then

\[
p_i(s_j|s_i) = \Pr_i(S_j = s_j | S_i = s_i)
\]

(1)

denote agent \( i \)’s signal posterior belief that agent \( j \) receives signal \( s_j \) given agent \( i \)’s signal \( s_i \). We refer to \( p_i(\cdot | \cdot) \) as agent \( i \)’s belief model. A crucial assumption for the
existence of strictly incentive compatible peer prediction mechanisms is that every agent’s belief model satisfies stochastic relevance [Johnson et al. 1990].

**Definition 2.1.** Random variable $S_i$ is *stochastically relevant* for random variable $S_j$ if and only if the distribution of $S_j$ conditional on $S_i$ is different for all possible values of $S_i$. That is, stochastic relevance holds if and only if $p_i(\cdot | s_i) \neq p_i(\cdot | s'_i)$ for all $i \in [n] := \{1, \ldots, n\}$ and all $s'_i \neq s_i$. Intuitively, one can think of stochastic relevance as correlation between different agents’ signal observations. Similar to the signal posteriors, we denote agent $i$’s signal prior belief about signal $s_j$ by $p_i(s_j) = \Pr_i(S_j = s_j)$. Note that for the signal prior and the signal posteriors it holds that $p_i(s_j) = \sum_{k=1}^{m} p_i(s_j | k) \cdot p_i(k)$.

### 2.2. Peer Prediction Mechanisms

We are now ready to define peer prediction mechanisms. For a discussion of possible extensions to more general mechanisms, see Section 8.

**Definition 2.2.** A (minimal) peer prediction mechanism is a function $M : [m] \times [m] \rightarrow \mathbb{R}$, where $M(x_i, x_j)$ specifies the payment to agent $i$ when she reports signal $x_i$ and her peer agent $j$ reports signal $x_j$.

We use ex post subjective equilibrium [Witkowski and Parkes 2012b], which is the most general solution concept for which truthful peer prediction mechanisms are known.

**Definition 2.3.** Mechanism $M$ is truthful if we have

$$s_i = \arg\max_{x_i} \mathbb{E}_{S_j} [M(x_i, S_j) \mid S_i = s_i]$$

for all $i \in [n]$ and all $s_i \in [m]$ with the expectation taken using agent $i$’s belief model, i.e. $S_j \sim p_i(\cdot | s_i)$.

The equilibrium is subjective because it allows for each agent to have a distinct belief model, and ex post because it allows for (but doesn’t require) knowledge of other agents’ belief models. Ex post subjective equilibrium is strictly more general than Bayes-Nash equilibrium (BNE) as it coincides with BNE when all agents share the same belief model, i.e. if $p_i(\cdot | \cdot) = p_j(\cdot | \cdot)$ for all $i, j \in [n]$.

**Definition 2.4.** A mechanism $M'$ is a positive-affine transformation of mechanism $M$ if there exists $f : [m] \rightarrow \mathbb{R}$ and $\alpha > 0$ such that for all $x_i, x_j \in [m]$, $M'(x_i, x_j) = \alpha M(x_i, x_j) + f(x_j)$.

The importance of Definition 2.4 lies in the fact that if $M$ is truthful, then $M'$ is truthful as well. As we will see, in certain cases these are the only possible truthful mechanisms.

**Lemma 2.5.** Let $M'$ be a positive-affine transformation of $M$. Then $M'$ is truthful if and only if $M$ is truthful.

**Proof.** The result follows directly from the fact that the $\arg\max$ from Definition 2.3 is unchanged by positive-affine transformations. That is, for all $i, j \in [n]$ and all $s_i, x_j \in [m]$. 

$$s_i = \arg\max_{x_i} \mathbb{E}_{S_j} [M(x_i, S_j) \mid S_i = s_i]$$

for all $i \in [n]$ and all $s_i \in [m]$ with the expectation taken using agent $i$’s belief model, i.e. $S_j \sim p_i(\cdot | s_i)$.
R. Frongillo and J. Witkowski

\[ s_i = \arg\max_{x_i} \mathbb{E}_{S_j} \left[ M(x_i, S_j) \mid S_i = s_i \right] \]
\[ = \arg\max_{x_i} \mathbb{E}_{S_j} \left[ \alpha M(x_i, S_j) + f(x_j) \mid S_i = s_i \right] \]
\[ = \arg\max_{x_i} \mathbb{E}_{S_j} \left[ M'(x_i, S_j) \mid S_i = s_i \right], \]

where the second line follows from the first because \( f(x_j) \) does not depend on \( x_i \).

2.3. The Probability Simplex

The intuition for our main results can be provided for \( m = 3 \) signals already, and so we give such examples throughout the paper. For probability distributions over only 3 signals, there is a convenient graphical representation of the probability simplex \( \Delta_m \) as an equilateral triangle, where the three corners represent the signals (see Figure 1L). The closer a point is to a corner (the distance from the corner’s opposing side), the more probability mass of that corner’s signal is on that point.\(^2\) The triangular shape ensures that for any point on the triangle the values of the three dimensions sum up to 1. For example, the point \( y(a, b, c) = (1/2, 1/3, 1/6) \) in Figure 1L is at height \( 1/6 \) (since the top corner represents signal \( c \)), and one half away from the right side of the triangle (because the left corner represents signal \( a \)). Observe that with three signals, there are only two degrees of freedom, and so fixing the point’s position with respect to \( a \) and \( c \), the value for signal \( b \) is fixed as well. (Confirm that \( y \) is one third away from the left side.)

2.4. Power Diagrams

Our results rely on a concept from computational geometry known as a power diagram, which is a type of weighted Voronoi diagram [Aurenhammer 1987b].

**Definition 2.6.** A power diagram is a partitioning of \( \Delta_m \) into sets called cells, defined by a collection of points \( \{v_s \in \mathbb{R}^m : s \in [m] \} \) called sites with associated weights \( w(s) \in \mathbb{R} \), given by

\[ \text{cell}(v^*) = \left\{ u \in \mathbb{R}^m : s = \arg\min_{x \in [m]} \{ \|u - v^x\|^2 - w(x) \} \right\}. \]

We call \( \|u - v^x\|^2 - w(x) \) the power distance from \( u \) to site \( v^x \); thus, for every point \( u \) in \( \text{cell}(v^*) \), it holds that \( v^x \) is closer to \( u \) in power distance than any other site \( v^x \).

We have defined power diagrams for the special case of the probability simplex, which is the case we need in this paper. The more general definition allows for a different number of sites than dimensions.\(^3\) The usual definition of a Voronoi diagram follows by setting all weights \( w(s) \) to 0.

3. MECHANISMS AND POWER DIAGRAMS

As with previous work, we would like to make statements of the form, “As long as the belief models satisfy certain constraints, the mechanism is truthful.” For example, the Shadowing Method [Witkowski and Parkes 2012a; Witkowski 2014] is truthful if and

\(^2\)This representation is equivalent to the natural embedding into \( \mathbb{R}^3 \) and viewing in the direction \((-1, -1, -1)\).

\(^3\)Also, note that we exclude cell boundaries; see Theorem 4.1.
only if \( p_i(s|s) - y(s) > p_i(s'|s) - y(s') \) for all \( s, s' \in [m] : s' \neq s \), and some distribution \( y(\cdot) \), which is a parameter of the mechanism (also see Figure 1L). When used directly (and not as a building block for more complex mechanisms), it is often assumed that there is a known, common signal prior \( p(\cdot) = p_i(\cdot) \) for all \( i \in [n] \), which is then used as \( y(\cdot) = p(\cdot) \). As we will see, both the Shadowing Method and the 1/p mechanism [Jurca and Faltings 2008; Jurca and Faltings 2011] are actually robust in that they are truthful even if there is no common signal prior. All that is required is that the agents’ possible posteriors fall into the correct regions. While it has been known that the constraints required by the Shadowing Method and the 1/p mechanism are incomparable, i.e. there exist belief models for which the Shadowing Method is truthful but the 1/p mechanism is not, and vice versa [Witkowski 2014], it was not known for which constraints there exist truthful mechanisms. In this section, we answer this question, and characterize all belief model constraints for which truthful mechanisms exist.

Our techniques derive heavily from the literature on property elicitation, where one wishes to extract a particular function, or property, of an agent’s belief using a scoring rule with access to a single sample from the true distribution (unlike our setting, where no ground truth is ever observed). Formally, a scoring rule \( S(\cdot, \cdot) \) elicits a property \( \Gamma \) if for all agent beliefs \( p \), the expected score \( \mathbb{E}_{x \sim p}[S(r, x)] \) is maximized by the report \( r = \Gamma(p) \). In particular, we leverage results from the finite property case, where the reports \( r \) are restricted to a finite set [Lambert and Shoham 2009]. For example, the mode of a distribution over \( m \) possible outcomes, \( \Gamma_{\text{mode}}(p) = \arg \max_s p(s) \), has \( m \) possible values. In peer prediction, these finite properties arise when examining the behavior of an agent facing a minimal peer prediction mechanism \( M \). Here, for an agent with posterior belief \( p \) about the reference agent’s signal, one could encode the agent’s optimal report into a function \( \Gamma_M \), where \( \Gamma_M(p) = s \) if \( s \) is the optimal report for \( p \) under \( M \). By definition then, the mechanism \( M \) can be thought of as a scoring rule eliciting \( \Gamma_M \). For example, the output agreement mechanism \( M(x_i, x_j) = \mathbb{1}\{x_i = x_j\} \) elicits the mode \( \Gamma_M = \Gamma_{\text{mode}} \). (Here \( \mathbb{1}\{\cdot\} \) is the indicator function.)

Of course, for \( M \) to be a truthful peer prediction mechanism, it must be the case that \( s \) is the optimal report for the posterior \( p(\cdot|s) \) following observation \( s \); this condition translates to \( \Gamma_M(p(\cdot|s)) = s \). One can now see how this corresponds to belief model constraints: a mechanism \( M \) is truthful with respect to some belief model \( p(\cdot|\cdot) \) if and only if \( \Gamma_M(p(\cdot|s)) = s \) for all signals \( s \). For example, output agreement is truthful if and only if \( s \) is the mode of \( p(\cdot|s) \) for all \( s \). At this point, we can use results in the property elicitation literature showing an equivalence between elicitable finite properties and power diagrams [Lambert and Shoham 2009; Frongillo and Kash 2014], first applying this equivalence to minimal peer prediction mechanisms, and then leverage techniques from computational geometry to show further structure, e.g. Theorems 4.1 and B.1.

As a first step, we formally define these constraints on belief models that limit which posteriors are possible following which signal.

Definition 3.1. A belief model constraint is a collection \( \mathcal{D} = \{D_s \subseteq \Delta_m : s \in [m]\} \) of disjoint sets \( D_s \) of distributions. If additionally we have \( \text{cl}(\cup_s D_s) = \Delta_m \), i.e. if \( \mathcal{D} \) partitions the simplex, we say \( \mathcal{D} \) is maximal.

A belief model constraint \( \mathcal{D} = \{D_1, \ldots, D_m\} \) ensures that for each agent \( i \), following signal observation \( S_i = s_i \), her belief about her peer agent’s signal \( s_j \) is restricted to be in \( D_{s_j} \). It is easy to come up with non-maximal belief model constraints, such as “\( \forall s \ p(s|s) > 0.6 \)” (Figure 3M). Note that under such a constraint, some distributions are not valid posteriors for any signal. In contrast, a maximal constraint covers the simplex, partitioning it into \( m \) bordering but non-overlapping regions (Figure 1L).

We can now talk about mechanisms being truthful with respect to belief model constraints.
We will now observe that for any mechanism $M$, there is a belief model constraint $D^M$, which exactly captures the set of belief models for which $M$ is truthful. In other words, not only is $M$ truthful with respect to $D^M$, but under any belief model that does not satisfy $D^M$, $M$ will not be truthful. The construction of $D^M$ is easy: for each signal $s$, $D^M_s$ is the set of distributions $p(\cdot | s)$ under which $x_i = s$ is the unique optimal report for $M$. Note that if $D^M_s$ is empty for any $s$, then $M$ is not truthful for any belief model.

**Lemma 3.3.** Let $M : [m] \times [m] \to \mathbb{R}$ be an arbitrary mechanism, and let $D^M$ be the belief model constraint given by

$$D^M_s = \left\{ p_i(\cdot | s) : s = \arg\max_{x_i} \mathbb{E}_{S_j \sim p_i(\cdot | s)} M(x_i, S_j) \right\}.$$
Then $M$ is truthful with respect to $D^M$, but not truthful for belief models not satisfying $D^M$. Moreover, if the rows of $M$ are all distinct, $D^M$ is maximal.

**Proof.** Suppose $p_i(\cdot|s) \in D^M_s$ for all $i \in [n], s \in [m]$. Then by construction of $D^M_s$, an agent $i$ receiving signal $s$ maximizes expected payoff by reporting $s$, and hence $M$ is truthful. By definition then, $M$ is truthful with respect to $D^M$. Now suppose $p_i(\cdot|s) \notin D^M_s$ for some $i \in [n], s \in [m]$. Then $s \neq \arg\max_{x_i} \mathbb{E}_{S_j \sim p} M(x_i, S_j)$, and thus $M$ cannot be truthful. Finally, consider the convex function $G(p) = \max_x \mathbb{E}_{S_j \sim p} M(x, S_j)$. By standard results in convex analysis (c.f. [Frongillo and Kash 2014, Theorem 3]) $G$ has subgradient $M(x, \cdot)$ whenever $x$ is in the argmax. As the rows of $M$ are all distinct, multiple elements in the argmax corresponds to multiple subgradients of $G$, and thus $G$ is nondifferentiable$^4$ at the set of indifference points $\{p : \arg\max_x \mathbb{E}_{S_j \sim p} M(x, S_j) \geq 2\}$. As $G$ must be differentiable almost everywhere [Aliprantis and Border 2007, Theorem 7.26], these indifference points must have measure 0 in the probability simplex, and thus $D^M$ is maximal. □

### 3.2. Equivalence to Power Diagrams

We have seen that every mechanism $M$ induces some belief model constraint $D^M$, and that $M$ is truthful with respect to $D^M$. We now show further that $D^M$ is a power diagram, and conversely, that every power diagram has a mechanism such that $D^M_s = \text{cell}(v^s)$ for all $s$.

The concrete mapping is as follows. Given mechanism $M : [m] \times [m] \to \mathbb{R}$, we construct sites and weights by:

$$v^s = M(s, \cdot), \quad w(s) = \|v^s\|^2 = \|M(s, \cdot)\|^2.$$  \hspace{1cm} (2)

Conversely, given a power diagram with sites $v^s$ and weights $w(s)$, we construct the mechanism $M$ as follows:

$$M(x_i, x_j) = v^{x_i}(x_j) - \frac{1}{2}\|v^{x_i}\|^2 + \frac{1}{2}w(x_i),$$  \hspace{1cm} (3)

where $v^{x_i}(x_j)$ is the $x_j$th entry of $v^{x_i}$. We note that these formulas are more explicit versions of those appearing in property elicitation, as mentioned above.

With these conversions in hand, we can now show that they indeed establish a correspondence between minimal peer prediction mechanisms and power diagrams.

**Theorem 3.4.** Given any mechanism $M : [m] \times [m] \to \mathbb{R}$, the induced belief model constraint $D^M$ is a power diagram. Conversely, for every power diagram given by sites $v^s$ and weights $w(s)$, there is a mechanism $M$ whose induced belief model constraint $D^M$ satisfies $D^M_s = \text{cell}(v^s)$ for all $s$.

**Proof.** Observe that if either relation (2) or (3) holds, we have the following for all $x, p$:

$$-2p \cdot v^x + \|v^x\|^2 - w(x) = -2 \mathbb{E}_{S_j \sim p} M(x, S_j).$$  \hspace{1cm} (4)

To see this, note that $p \cdot M(x, \cdot) = \mathbb{E}_{S_j \sim p} M(x, S_j)$. Adding $\|p\|^2$ to both sides of Eq. 4 gives

$$\|p - v^x\|^2 - w(x) = \|p\|^2 - 2 \mathbb{E}_{S_j \sim p} M(x, S_j).$$  \hspace{1cm} (5)

$^4$Technically, we should restrict to the first $m - 1$ coordinates of the distribution, or use the approach of Appendix A, so that $G$ is defined on a full-dimensional subset of $\mathbb{R}^{m-1}$. Neither alters the argument.
Now applying Eq. 5 to the definitions of a power diagram and of $D^M$, we have
\[ p \in \text{cell}(v^s) \iff s = \arg\min_x \{ \|p - v^x\|^2 - w(x) \} \]
\[ \iff s = \arg\min_x \left\{ \|p\|^2 - 2 \mathbb{E}_{S_j \sim p} [M(x, S_j)] \right\} \]
\[ \iff s = \arg\max_x \mathbb{E}_{S_j \sim p(\cdot|s)} M(x, S_j) \]
\[ \iff p \in D^M_s . \]

Finally, as Eq. 2 defines a power diagram for any mechanism $M$, and Eq. 3 defines a mechanism for any power diagram, we have established our equivalence.

**Corollary 3.5.** Let $\mathcal{D}$ be a maximal belief model constraint. Then there exists a mechanism that is truthful with respect to $\mathcal{D}$ if and only if $\mathcal{D}$ is a power diagram.

We give illustrations of Corollary 3.5 in Figure 1. In particular, Figure 1L shows a maximal belief model constraint that is a power diagram, and thus there must be a truthful mechanism (in this case, the Shadowing Method). The other two figures are negative examples, either because the constraint is maximal but not a power diagram (Figure 1M) or because even though the constraint is non-maximal, there is no consistent power diagram (Figure 1R). These examples illustrate the power of this geometric approach in determining whether truthful mechanisms exist, even for non-maximal constraints.

A surprising corollary of Theorem 3.4 comes when connecting the bijection result to a class of diagrams known as Bregman Voronoi diagrams, which are Voronoi diagrams with the distance given by a (typically asymmetric) Bregman divergence in- stead of Euclidean distance [Nielsen et al. 2007; Frongillo and Kash 2014]. Bregman divergences are known to be equivalent to proper scoring rules [Gneiting and Raftery 2007], making Bregman Voronoi diagrams precisely those which arise from classical peer prediction mechanisms. Somewhat surprisingly, Nielsen [2007] show that Bregman Voronoi diagrams are equivalent to power diagrams. Applying Theorem 3.4, we have minimal PP $\iff$ power diagram $\iff$ Bregman Voronoi $\iff$ classical PP, or in other words, every minimal mechanism can be written as a classical peer prediction mechanism with respect to some proper scoring rule.

**Corollary 3.6.** Every minimal peer prediction mechanism $M$ can be written as a classical peer prediction mechanism $M(x_i, x_j) = S(p^x_i, x_j)$ for some choice of proper scoring rule $S$ and sites $\{p^s\}_{s \in [m]}$ in the affine hull of $\Delta_m$.

Note that the sites in Corollary 3.6 need not be in the simplex. Also, while it is easy to construct such a proper scoring rule $S$ given $M$ by expressing it as the Bregman divergence with respect to the convex function $G(p) = \arg\max_{x_i, x_j} \mathbb{E}_{x_i, x_j \sim p} [M(x_i, x_j)]$, the resulting score $S$ will not be strictly proper, as $G$ is not strictly convex. It seems intuitively clear that some other $S$ can be chosen to be strictly proper, yet an explicit construction remains an open question. Such a construction would involve finding a strictly convex function $G$ and points $p^s$ satisfying $v^s = \frac{1}{2} \nabla G(p^s)$ and $w(s) = \frac{1}{4} \|\nabla G(p^s)\|^2 + G(p^s) - p^s \cdot \nabla G(p^s)$ [Frongillo and Kash 2014, Appendix B].

Finally, it is easy to see that the conversion from mechanisms to power diagrams (Eq. 2) and back (Eq. 3) are inverse operations. In particular, this shows that mechanisms are in one-to-one correspondence with power diagrams on $\Delta_m$. In the following

---

5A Bregman divergence is given by $D_G(u, v) = G(u) - G(v) - \nabla G(v) \cdot (u - v)$ for some convex function $G$. 

section, we will leverage this tight connection, and use results from computational geometry to show that several well-known mechanisms are unique in the sense that they are the only mechanisms, up to positive-affine transformations, that are truthful for their respective belief model constraints.

4. UNIQUENESS

Consider the standard output agreement mechanism \( M(x_i, x_j) = 1 \) if \( x_j = x_i \) and 0 otherwise. It is easy to see that the mechanism is truthful as long as each agent assigns the highest posterior probability \( p_i(\cdot | s_i) \) to their own signal \( s_i \), yielding the constraint \( \forall s' \neq s \). The type of question we address in this section is: are there any other mechanisms than \( M \) that are guaranteed to be truthful as long as posteriors satisfy this condition? We will show that, up to positive-affine transformations, the answer is no: output agreement is unique. Moreover, the Shadowing Method and the \( 1/p \) mechanism are also unique for their respective conditions on posteriors.

To get some intuition for this result, let us see why output agreement is unique for \( m = 3 \) signals \( a, b, c \). From Theorem 3.4, we know that \( D = D^M \), the induced belief model constraint, is a power diagram (which is depicted in Figure 2L). In general, there may be many sites and weights that lead to the same power diagram, and these may yield different mechanisms via Eq. 3. In fact, it is a general result that any positive scaling of the sites followed by a translation (i.e., some \( \alpha > 0 \) and \( u \in \mathbb{R}^m \) so that \( \bar{v}^s = \alpha v^s + u \) for all \( s \)) will result in the same power diagram for an appropriate choice of weights [Aurenhammer 1987b]. As it turns out, such scalings and translations exactly correspond to positive-affine transformations when passing to a mechanism through Eq. 3. Thus, we only need to show that the sites for the output agreement power diagram are unique up to scaling and translation. Here, another useful property of sites comes into play: the line between sites of two adjacent cells must be perpendicular to the boundary between the cells. Examining Figure 2L, one sees that after fixing \( v^a \), the choice of \( v^b \) is constrained to be along the blue dotted line, and once \( v^a \) and \( v^b \) are chosen, \( v^c \) is fixed as the intersection of the red dotted lines. Thus, we can specify the sites by choosing \( v^a \) (a translation), and how far away from \( v^a \) to place \( v^b \) (a positive scaling). We can then conclude that output agreement for three signals is unique up to positive-affine transformations. We now give the general result.

**Theorem 4.1.** If there exists a mechanism \( M \) that is truthful for some maximal belief model constraint \( D \), and there is some \( y \in \Delta_m \) with \( y(s) > 0 \) \( \forall s \) such that \( \cap_s \text{cl}(D_s) = \{y\} \), then \( M \) is the unique truthful mechanism for \( D \) up to positive-affine transformations.
PROOF. As $M$ is truthful with respect to $D$, we have $D = D^M$ and thus $D$ is a power diagram $P$ from Theorem 3.4. By our assumption, we observe that the only vertex (also called a 0-dimensional face) of $P$ must be $y$, the intersection of the $\text{cl}(D_y)$, as there are $m$ cells but $\Delta_m$ has dimension $m - 1$. Throughout the proof, as before, we implicitly work in the affine hull $A$ of $\Delta_m$. We may assume without loss of generality that all sites lie in $A$, as we may translate any site to $A$ while adjusting its weight to preserve the diagram and resulting mechanism (see Appendix A). Thus, by definition of a simple cell complex, as the vertex $y$ is in the relative interior of $\Delta_m$, we see that the extension $\hat{P}$ of $P$ onto the affine hull of $\Delta_m$ must also have $y$ as the only vertex. (As a vertex is on the boundary of every cell, and there are $m$ cells in $m - 1$ dimensions with disjoint interiors, there can be only one such point.) Now following the proof of Frongillo and Kash [2014, Theorem 4], we note that Aurenhammer [1987b, Lemma 1] and Aurenhammer [1987a, Lemma 4] together imply the following: if the proof of Frongillo and Kash [2014, Theorem 4], we observe that different choices of sites $\{\hat{v}^s\}_{s \in [m]}$ and weights $w(\cdot)$, then any other representation of $\hat{P}$ with sites $\{\hat{v}^s\}_{s \in [m]}$ satisfies $\exists \alpha > 0, u \in \mathbb{R}^m$ s.t. $\hat{v}^s = \alpha v^s + u$ for all $s \in [m]$. In other words, all sites must be a translation and scaling of $\{v^s\}$. To complete the proof, we observe that different choices of $u$ and $\alpha$ (with suitable weights) merely yield an affine transformation of $M$ when passed through Eq. 3, and as any positive-affine transformation preserves truthfulness, the result follows. 

COROLLARY 4.2. The following mechanisms are unique, up to positive-affine transformations, with respect to the corresponding constraints (each with “$v^s' \neq s^s$ and “$v_i \in [n]$” implied):

1. Output Agreement, $p_i(s|s) > p_i(s'|s)$
2. Shadowing Method, $p_i(s|s) - y(s) > p_i(s'|s) - y(s')$
3. $1/p$ Mechanism, $p_i(s|s)/y(s) > p_i(s'|s)/y(s')$.

PROOF. In all three cases, the given mechanism is known to be truthful for its respective belief model constraint. Moreover, for all three constraints $D$, one can check that $\cup_x \text{cl}(D_x) = \Delta_m$ and $\cap_x \text{cl}(D_x) = \{y\}$ meaning that $y$ is the unique distribution bordering every set $D_x$ (for Output Agreement, $y$ is the uniform distribution). Hence, the mechanisms are unique up to positive-affine transformations by Theorem 4.1. 

Let us make a few remarks on Theorem 4.1. First, note that the restriction that $y$ must not touch the boundary of the simplex is necessary, as uniqueness need not (will not) hold otherwise; see Figure 3L. Similarly, for constraints $D$ that are not maximal, there may be many more truthful mechanisms; Figure 3M depicts two distinct power diagrams yielding mechanisms that are truthful with respect to the non-maximal constraint “$p(s|s) > 0.6$”, and thus they are not merely positive-affine transformations of each other. That said, some non-maximal constraints $D$ still yield a unique truthful mechanism, as illustrated in Figure 3R.

In Appendix B, we strengthen the results of this section, showing that under the same conditions as Theorem 4.1, not only is the mechanism unique up to positive-affine transformations, but it can be expressed as a positive-affine transformation of a classical peer prediction mechanism with respect to the Brier score. This is done by first showing that in this setting, the maximal constraint $D$ is not just a power diagram but in fact a Voronoi diagram.

---

*A cell complex $P$ is simple if each of $P$’s vertices is a vertex of exactly $m$ cells of $P$, the minimum possible [Aurenhammer 1987a, p.50].*
5. OPTIMAL MECHANISMS FOR NEW CONDITIONS

In this section, we show how to compute new mechanisms that are truthful with respect to new conditions. Moreover, we find the positive-affine transformation that maximizes effort incentives subject to a budget. It then follows directly from Theorem 4.1 that the final mechanism’s effort incentives are globally optimal given this condition. That is, there is no peer prediction mechanism that is truthful with respect to the new condition providing better effort incentives.

5.1. Computing Truthful Mechanisms

To construct mechanisms from belief model constraints, we turn to computational geometry to find sites and weights for the corresponding power diagram. For the class of constraints satisfying Theorem 4.1, where there is an intersection point $y$ in the interior of the simplex, which is the most common case in peer prediction, one can rely on the $O(m)$ time algorithm given by Aurenhammer [1987c].

We now give a version of this procedure, specialized for our setting.

The first consideration is the form of the input to the procedure. How should we assume the belief model constraint $D$ is given? One succinct representation is an $m \times m$ matrix representing a truthful mechanism for the constraint, but of course specifying a constraint with a mechanism defeats the whole purpose of the algorithm. To be useful, we need a representation for the belief model constraint that is closer to the partitioning point of view. Here, we assume that one knows the constraint boundary between any pair of signals $(s, s')$. If $D$ is a power diagram (a necessary condition for the existence of a truthful mechanism by Theorem 3.4) this boundary must be defined by a hyperplane, which in turn can be represented by its defining normal vector $u^{s,s'}$. All that remains is specifying the orientation of $u^{s,s'}$ so we know which direction is which cell; here we assume $u^{s,s'}$ points in the direction away from $D_s$, so that we have $p \in D_s \iff u^{s,s'} \cdot p < 0$ for all $s' \neq s$. As an example, the shadowing constraint is given by $u^{s,s'} = 1_{s'} - 1_s - (y(s') - y(s))1$, where $1_s$ is the standard unit vector (with 1 in coordinate $s$ and 0 otherwise), and $1$ is the all-ones vector. Hence,

---

Footnote 7: This condition is a special case of the so-called simple power diagrams; see footnote 6. For more complicated cases, one would use [Rybnikov 1999].
\[
\mathbf{u}^{s,s'} \cdot \mathbf{p} < 0 \iff p(s') - p(s) - y(s') + y(s) < 0, \quad \text{since} \quad 1 \cdot p = 1.
\]
For the condition used in Section 6, one can take \(\mathbf{u}^{s,s'} = (1 - y(s)) \mathbf{I}_s - (1 - y(s')) \mathbf{I}_{s'} + (y(s) - y(s')) \mathbf{I}.

Thus, given any maximal belief model constraint \(\mathcal{D}\), we can represent \(\mathcal{D}\) by a collection of vectors \(\{\mathbf{u}^{s,s'} \in \mathbb{R}^m : s, s' \in [m]\}\). With this representation in hand, the following algorithm computes a truthful mechanism for \(\mathcal{D}\), under the additional assumption that \(\mathcal{D}\) satisfies the conditions of Theorem 4.1.

**Algorithm 1** Compute Mechanism from Belief Model Constraint

i. Solve for \(y\) as the unique solution to \(\mathbf{u}^{s,s'} \cdot y = 0 \quad \forall s, s' \) and \(y \cdot \mathbf{1} = 1\).

ii. Let \(\hat{\mathbf{u}}^{s,s'} = \mathbf{u}^{s,s'} - \left(\frac{1}{m} \mathbf{u}^{s,s'} \cdot \mathbf{I}\right) \mathbf{1}\) for all \(s, s'\). \# Projects \(\mathbf{u}^{s,s'}\) onto the affine hull of the simplex \(\Delta_m\).

1: Choose \(s \in [m]\) and any \(v^s\) in the affine hull of \(\Delta_m\).
2: Choose \(s' \neq s\) and any \(\alpha > 0\), and set \(v^{s'} = v^s + \alpha \hat{\mathbf{u}}^{s,s'}\).
3: For all \(s'' \notin \{s, s'\}\), find the unique positive solution \((\beta, \gamma)\) to \(\alpha \hat{\mathbf{u}}^{s,s'} + \beta \hat{\mathbf{u}}^{s',s''} = \gamma \hat{\mathbf{u}}^{s,s''}\), and set \(v^{s''} = v^s + \gamma \hat{\mathbf{u}}^{s,s''}\). \# See argument below.
4: Set \(w(s) = 0\) and \(w(s'') = \|y - v^{s''}\|^2 - \|y - v^s\|^2\) for all \(s'' \neq s\).
5: Compute the mechanism by applying Eq. 3.

For the correctness of Algorithm 1, first observe that line i has a unique solution by the assumption of Theorem 4.1. We must also show uniqueness in line 3. This follows by the result of Theorem 4.1, which shows that two different collections of sites representing these constraints must be related by a positive-affine transformation. As we show now, this implies that once we have fixed \(v^s\) and \(v^{s'}\), the rest of the sites are uniquely determined.\(^8\) Consider some positive-affine transformation of the sites that keeps \(v^s\) and \(v^{s'}\). As \(v^s\) has not moved, the translation must be 0, and as \(v^{s'}\) also remains the same, so does the length of the line segment between \(v^s\) and \(v^{s''}\), so the scaling must be 1. We have now determined the positive-affine transformation: the identity. Thus all other sites are uniquely determined by the first two. Finally, one can check that the projections \(\hat{\mathbf{u}}^{s,s'}\) are correctly oriented and remain perpendicular to the boundary of cells \(D_s\) and \(D_{s'}\).\(^9\)

### 5.2. Optimizing Effort Incentives

From Section 5.1, we know how to compute a mechanism that is truthful with respect to a given belief model constraint. In this section, we take this one step further and optimize within the space of truthful mechanisms. As explained in Section 1, peer prediction mechanisms are especially useful for incentivizing effort, i.e. the costly acquisition of signals, and we will thus address the following optimization problem:

\[
\begin{align*}
\max & \quad \text{effort incentives} \\
\text{s.t.} & \quad \text{truthfulness} \\
& \quad \text{budget constraint} \\
& \quad \text{non-negative payments}
\end{align*}
\]

\(\max e_i(M) \quad \text{with respect to} \quad \mathcal{D} \quad M(x_i, x_j) \leq B \quad \text{M}(x_i, x_j) \geq 0 \quad (6)

---

\(^8\) Note that we are restricting the sites to the affine hull of the simplex, which is simply given by the condition \(v \cdot \mathbf{1} = 1\); see Appendix A. Without this, there would be an extra degree of freedom in choosing the sites, but as that section shows, this is irrelevant.

\(^9\) Letting \(p, q \in \Delta_m\) be distinct points on the boundary of these cells, and letting \(z = p - q\), we have \(\hat{\mathbf{u}}^{s,s'} \cdot z = \mathbf{u}^{s,s'} \cdot z - \left(\frac{1}{m} \mathbf{u}^{s,s'} \cdot \mathbf{1}\right) \cdot z = \mathbf{u}^{s,s'} \cdot z\) as \(p \cdot \mathbf{1} = q \cdot \mathbf{1} = 1\), so \(z = 0\).
Effort can be modeled in many different ways. Following Witkowski [2014], our modeling of effort is binary: agents either exert effort or not.

**Definition 5.1.** Given that agent $j$ invests effort and reports truthfully, the effort incentive $e_i(M)$ that is implemented for agent $i$ by peer prediction mechanism $M$ is the difference in expected utility of investing effort followed by truthful reporting and not investing effort, i.e.

$$e_i(M) = \mathbb{E}_{S_i, S_j} [M(S_i, S_j)] - \max_{x_i \in [m]} \mathbb{E}_{S_j} [M(x_i, S_j)].$$

where $x_i$ is agent $i$’s signal report that maximizes her expected utility according to the signal prior, and where the expectation is using agent $i$’s subjective belief model $p_i(\cdot | \cdot)$.

Thus, the effort incentive $e_i(M)$ is agent $i$’s expected gain by exerting effort. Naturally, scaling a mechanism should scale the incentives; in fact, this can be generalized to positive-affine transformations.

**Lemma 5.2.** For any mechanism $M$, and any positive-affine transformation $M' = \alpha M + f$, we have $e_i(M') = \alpha e_i(M)$.

**Proof.** This follows from a simple computation:

$$e_i(M') = \mathbb{E}_{S_i, S_j} [M'(S_i, S_j)] - \max_{x_i \in [m]} \mathbb{E}_{S_j} [M'(x_i, S_j)]$$

$$= \mathbb{E}_{S_i, S_j} [\alpha M(S_i, S_j) + f(S_j)] - \max_{x_i \in [m]} \mathbb{E}_{S_j} [\alpha M(x_i, S_j) + f(S_j)]$$

$$= \alpha \mathbb{E}_{S_i, S_j} [M(S_i, S_j)] + \mathbb{E}_{S_j} [f(S_j)] - \alpha \max_{x_i \in [m]} \mathbb{E}_{S_j} [M(x_i, S_j)] - \mathbb{E}_{S_j} [f(S_j)]$$

$$= \alpha e_i(M).$$

\(\Box\)

From Section 4 we know that the space to optimize over is restricted to positive-affine transformations of any truthful mechanism once the belief model constraint is fixed and given the conditions of Theorem 4.1. Using this, we can pin down the effort-maximizing mechanism as given by the following theorem.

**Theorem 5.3.** Let mechanism $M$ and belief model constraint $D$ satisfy the conditions of Theorem 4.1. Let $g(x_j) = \min_{x_j} M(x_i, x_j)$ and $\alpha = B/\max_{x_i, x_j} M(x_i, x_j) - g(x_j)$. Then mechanism $M' = \alpha M(x_i, x_j) - \alpha g(x_j)$ optimizes effort in Eq. 6.

**Proof.** We will show something slightly stronger, giving the full characterization of mechanisms optimizing Eq. 6. By Theorem 4.1, $M$ is the unique truthful mechanism for $D$ up to affine transformations $M_{\alpha, f}(x_i, x_j) = \alpha M(x_i, x_j) + f(x_j)$, so we need only search for the optimal $\alpha, f$. By Lemma 5.2, $e_i(M_{\alpha, f}) = \alpha e_i(M)$, which since $e_i(M)$ is a positive constant reduces the optimization to the following:

$$\max \alpha \quad \text{s.t.} \quad \exists f \forall x_i, x_j \; \alpha M(x_i, x_j) + f(x_j) \in [0, B].$$

Clearly then, the maximum value of $\alpha$ we could hope for is $\alpha = B/\max_{x_i, x_j} (\max_{x_i} M(x_i, x_j) - \min_{x_j} M(x_i, x_j))$, and indeed taking this $\alpha$, one sees that any $f$ in the range $f(x_j) \in [-\alpha \min_{x_i} M(x_i, x_j), -\alpha (\max_{x_i} M(x_i, x_j) - \min_{x_i} M(x_i, x_j))] \neq \emptyset$ will satisfy the constraints. One can check that in particular taking $f(x_j) =$
Fig. 4: The truthful mechanism with respect to the “complement 1/p” condition in power diagram form with sites $v^a$, $v^b$, and $v^c$. Notice that while the intersection point $y = (1/2, 1/3, 1/6)$ is the same as in Figure 1L, the belief model constraint as depicted by the dashed partitioning is now given by the new “complement 1/p” condition. (Compare to Figure 2.)

$-\alpha \min_{x_i} M(x_i, x_j)$ gives the mechanism in the theorem statement, which gives the lowest possible payments among the optimal mechanisms. As $f = -\alpha g$, we are done.

It is important to remark that even though the magnitude of an agent’s effort incentive will depend on the particular values of their posteriors $\{p_i(\cdot|s)\}_{s \in [m]}$, Theorem 5.3 maximizes the effort incentives of all agents regardless of their posteriors, as long as they satisfy the constraint $D$.

To illustrate the power of Theorem 5.3, consider the Shadowing Method [Witkowski and Parkes 2012a; Witkowski 2014], which has a parameter $\delta$ specifying how much to perturb $y$ in order to obtain the shadow posterior. A natural question to ask is for which $\delta$ the effort is maximized subject to the constraints in Eq. 6. A direct analysis is quite tedious, with many lines of algebra, and even with the optimal $\delta$ in hand, it is not clear whether one could do better by perhaps adding a score $f(x_j)$ that depends only on the peer agent’s signal report $x_j$, or by allowing for shadow posteriors that aren’t valid distributions followed by renormalizing the scoring rule so that the score is again bounded by $[0, B]$. Moreover, the optimal mechanism could have been of a different form entirely. Theorem 5.3 suggests a better approach to this problem: take the Shadowing Method with any $\delta > 0$ as a black box and simply find the maximum scaling allowing a translation that keeps the scores in the interval $[0, B]$.

6. EXAMPLE

In this section, we exemplify Algorithm 1 from Section 5, constructing a new mechanism that is truthful with respect to a new condition. Moreover, we compute the optimal such mechanism with respect to the effort it incentivizes as explained in Section 5.2.

As intuition for the new condition, imagine the mechanism has an estimate of the agents’ signal priors $p(a, b, c) = (0.01, 0.04, 0.95)$, which it designates as the intersection point $y(\cdot) = p(\cdot)$ of the belief model constraint. Consider now posterior $p_i(a, b, c|s) = (0.02, 0.01, 0.97)$, where the 1/p mechanism would pick signal $a$ since its relative increase from prior (as estimated by the mechanism) to posterior is highest (it doubles). However, one could also consider the relative decrease in “error”: in a world without noise, the posterior would have $p_i(s|s) = 1$ for every $s$, and so signal $a$’s relative decrease from $0.99 = 1 - 0.01$ to $0.98 = 1 - 0.02$ is not as “impressive” as signal $c$’s decrease in error from $0.05 = 1 - 0.95$ to $0.03 = 1 - 0.97$ (a reduction of almost one half).
Formalizing this intuition yields the “complement 1/p” condition, $\frac{1-y(s)}{1-p_i(s|s)} > \frac{1-y(s')}{1-p_i(s'|s)}$, $\forall s' \neq s$.

Theorem 4.1 implies that there is a unique mechanism that is truthful for this new condition, up to positive-affine transformations. We now exemplify the construction of the new “complement 1/p” condition following the steps of Algorithm 1. For that purpose, we return to our running example with $m = 3$ signals and intersection point $y = (1/2, 1/3, 1/6)$ as depicted in Figure 4.

1: Pick any point for $v^a$, say $v^a = (4/5, 1/10, 1/10)$.\(^\text{10}\)
2: Pick any $v^b$ on the blue dotted line, ensuring that the line between $v^a$ and $v^b$ is perpendicular to the $a,b$ cell boundary. Here we choose $v^b = (1/10, 81/110, 9/55)$.
3: For all other signals $s$, $v^a$ is now uniquely determined by $v^a$ and $v^b$ as the lines between any two sites must be perpendicular to their cell boundary. Here we only have one other signal, $c$, so we take $v^c$ to be the unique point at the intersection of the red dotted lines, which is $v^c = (38/275, 111/550, 33/50)$.
4: Calculate the weights by observing that $y$ must be equidistant (in the power distance) to all sites simultaneously: $w(a) = 0$, $w(b) = 23/100$, $w(c) = 548/1875$.
5: We obtain the resulting mechanism by applying Eq. 3:

$$M(\cdot, \cdot) = \frac{1}{1100} \begin{bmatrix} 517 & -253 & -253 \\ -113 & 587 & -43 \\ 13 & 83 & 587 \end{bmatrix}.$$

6: From Theorem 5.3, it then follows that, among all positive-affine transformations of $M$, the mechanism $M^*$ optimizing effort incentives given a budget $B = 1$ is:

$$M^*(\cdot, \cdot) = \frac{1}{20} \begin{bmatrix} 15 & 0 & 0 \\ 0 & 20 & 5 \\ 3 & 8 & 20 \end{bmatrix}.$$

This step can be computed as follows: subtract the largest amount from each column that keeps payments nonnegative, and then scale so the largest entry is 1.

We conclude by noting that this example condition admits a closed form solution: $M(s, s') = 0$ if $s = s'$, and $\frac{1}{1-w(s)}$ otherwise. One can check that adding constants to each column to give non-negative payments recovers $M^*$. Of course, our construction applies even when no convenient closed-form solution exists, such as in Figure 5R.

7. MAXIMALLY-ROBUST MECHANISMS

In the classical peer prediction method [Miller et al. 2005], the mechanism is assumed to have full knowledge of the agents’ belief models. Recent work relaxes the method’s knowledge requirements, e.g. using additional reports [Prelec 2004; Witkowski and Parkes 2012b; Witkowski and Parkes 2012a] or using reports on several items [Dasgupta and Ghosh 2013; Witkowski and Parkes 2013]. An approach closer to the classical method has been suggested by Jurca and Faltings [2007], who compute a minimal mechanism as the solution of a conic optimization problem that ensures truthfulness as long as the agents’ belief models are close to the mechanism’s, with respect to Euclidean distance. This restriction, a form of robustness, is defined as follows.

**Definition 7.1.** [Jurca and Faltings 2007] A mechanism $M$ is $\epsilon$-robust with respect to belief model $p(\cdot|\cdot)$ if $M$ is truthful for $p^*(\cdot|\cdot)$ whenever the following holds for all

\(^{10}\)While we chose $v^a \in D_a$, it could be in any other cell, e.g. $D_c$, or even outside the simplex.
\[ s_i \in [m], \]
\[ \sum_{s_i \in [m]} (p(s_j|s_i) - p^*(s_j|s_i))^2 \leq \epsilon^2. \quad (7) \]

While Jurca and Faltings fix the robustness \( \epsilon \) as a hard constraint, one may also seek the mechanism that maximizes this robustness. The achievable robustness is of course limited by the mechanism’s belief model \( p(\cdot|\cdot) \); in particular, the “robustness areas” around the mechanism’s posteriors cannot overlap; see Figure 5L. Viewing robustness in geometric terms, we obtain a closed-form solution.\(^{11}\)

**Theorem 7.2.** Let \( p(\cdot|\cdot) \) be the mechanism’s belief model in classical peer prediction. Then the following mechanism is maximally robust:

\[ M(x_i, x_j) = p(x_j|x_i) - \frac{1}{2} \sum_{s=1}^{m} p(s|x_i)^2. \quad (8) \]

**Proof.** In light of Theorem 3.4, we may focus instead on power diagrams. From Eq. 7, for all \( s \) we must have \( B_p(p(\cdot|s)) \subseteq \text{cell}(v^*) \), where \( B_p(u) \) is the Euclidean ball of radius \( \epsilon \) about \( u \) (restricted to the probability simplex). Letting \( d = \min_{s,s' \in [m]} \| p(\cdot|s) - p(\cdot|s') \| \) be the minimum Euclidean distance between any two posteriors, it becomes clear that robustness of \( d/2 \) or greater cannot be achieved, as \( \frac{1}{2} \| p(\cdot|s) - p(\cdot|s') \| \) be the minimum Euclidean distance between any two posteriors, it becomes

\[ \frac{1}{2} \| p(\cdot|s) - p(\cdot|s') \| \text{ for all } s \in [m]. \]

**Corollary 7.3.** The classical peer prediction method with the quadratic scoring rule is maximally robust.

One can easily adapt the above to design maximally robust mechanisms with respect to non-Euclidean distances as well, so long as that distance can be expressed as a Bregman divergence (see footnote 5 and Corollary 3.6). Each such divergence has a corresponding scoring rule which one simply uses in the place of the quadratic score [Frongillo and Kash 2014, Appendix F].

**8. DISCUSSION AND CONCLUSION**

We have presented a new geometric perspective on minimal peer prediction mechanisms, and proved that it is without loss of generality to think of a minimal peer prediction mechanism as a power diagram. This perspective then allowed us to prove uniqueness of several well-known mechanisms up to positive-affine transformations, to construct novel peer prediction mechanisms for new conditions, to optimize for effort incentives within this space, and to compute the mechanism that is maximally robust with respect to the agents’ subjective belief models deviating from the center’s.

Several extensions of our model and results are straightforward. For example, mechanisms that score an agent with the reports of multiple reference agents are still equivalent to power diagrams, but on a higher-dimensional belief space (all distributions on pairs of signals); uniqueness, however, would fail to hold in general. Alternatively,
if agents are each given separate mechanisms $M_i(x_i, x_j)$, essentially all of our main results go through, including uniqueness of each $M_i$ with respect to a belief model constraint $D_i$.

We believe the most exciting direction for future work is to construct mechanisms from real-world data. One way to do this (aside, of course, from using gold standard data), is to use the geometric framework to learn Bayesian Truth Serum mechanisms from the agents’ reports. In addition to the signal report, these mechanisms also elicit posterior reports, and with these (signal, posterior) pairs in hand, the mechanism designer can then train a classifier within the class of power diagrams that predicts the signal associated with a new posterior. (Note that multi-class Support Vector Machines naturally produce power diagrams [e.g. Borgwardt 2015].) This power diagram can then be converted to a mechanism using Eq. 3. If a max-margin criterion is imposed when training, as depicted in Figure 5R, the resulting mechanism will be maximally robust with respect to the training set. When the data are not linearly separable, a soft-margin solution may be appropriate. We explore this approach in ongoing work.

**ACKNOWLEDGMENTS**

We thank anonymous reviewers for very useful comments and feedback.

**REFERENCES**


A. WORKING IN THE AFFINE HULL OF THE SIMPLEX

In this section we will justify the step in Sections 3 and 4 where we assume without loss of generality that we may work with power diagrams in \( d = m - 1 \) dimensions, where as always \( m \) is the number of signal values. Specifically, we show that we may assume WLOG that all sites \( v^* \) are in the affine hull \( A = \{ u \in \mathbb{R}^m : \sum_u u(s) = 1 \} \) of \( \Delta_m \). Then, as this affine subspace has dimension \( d \), we can apply an isometry to restate the problem in \( \mathbb{R}^d \) (which exists as they are both flats of the same dimension), so that we can work with \( \hat{v}^1, \ldots, \hat{v}^m, \hat{p} \in \mathbb{R}^d \).

To begin, we show that assuming \( v^* \in A \) is WLOG. Let \( p \in \Delta_m \) and \( v^* \in \mathbb{R}^m \), and define \( \hat{v}^* = v^* + c(s) \mathbb{1} \) for \( c(s) \in \mathbb{R} \) where \( \mathbb{1} \in \mathbb{R}^m \) is the all-ones vector. Then we have

\[
\| \hat{v}^* - p \|^2 = \| v^* + c(s) \mathbb{1} \|^2 - 2( v^* + c(s) \mathbb{1} ) \cdot p + \| p \|^2 = \| v^* \|^2 + 2v^* \cdot c(s) \mathbb{1} + \| c(s) \mathbb{1} \|^2 - 2v^* \cdot p - 2c(s) + \| p \|^2 = \| v^* - p \|^2 + c(s)^2 m + 2c(s)(v^* \cdot \mathbb{1} - 1)
\]

Taking \( c(s) = (1 - v^* \cdot \mathbb{1})/m \), we have \( \hat{v}^* \cdot \mathbb{1} = 1 \) so \( \hat{v}^* \in A \). We now check that taking \( \hat{w}(s) = w(s) - (1 - v^* \cdot \mathbb{1})^2/m \) preserves power distances. From the above, we see that

\[
\| \hat{v}^* - p \|^2 - \hat{w}(s) = \| v^* - p \|^2 - w(s) + \frac{(1 - v^* \cdot \mathbb{1})^2}{m} + \frac{(1 - v^* \cdot \mathbb{1})^2}{m} m + 2 \frac{(1 - v^* \cdot \mathbb{1})}{m} (v^* \cdot \mathbb{1} - 1) = \| v^* - p \|^2 - w(s)
\]
A Geometric Method to Construct Minimal Peer Prediction Mechanisms

Fig. 6: (L) A power diagram in the plane with 4 sites that is not a Voronoi diagram. If the sites are the vertices of the unit square, the weights are \( w(a) = w(d) = 0 \) and \( w(b) = w(c) = 1/2 \). To see that the diagram cannot be Voronoi, note that because the diagram is simple (see footnote 6) the sites drawn are unique up to positive-affine transformations; due to the angles of the cell boundaries, they must always form a square with the correct orientation. But clearly the Voronoi diagram for such sites would always be "+" shaped, so this diagram cannot be Voronoi. Note that with 3 sites, every power diagram is actually Voronoi: either the cell boundaries are parallel or they all intersect at some point \( y \), the former of which is trivially Voronoi and the latter of which is Voronoi by Theorem B.1.

(R) For an example in our peer prediction setting, we must consider \( m = 4 \) signals to have four sites, where the simplex \( \Delta_4 \) can be visualized as a tetrahedron. Here we simply embed the sites on a plane within the tetrahedron in the same way; for example, one could take sites \( v^a = (0.1, 0.1, 0.7) \), \( v^b = (0.4, 0.1, 0.3), v^c = (0.1, 0.4, 0.1, 0.3), v^d = (0.4, 0.4, 0.1, 0.1) \), where the first two coordinates are tracing out a square on the plane \( \{ v : v_3 = 0.1 \} \) shown as dashed lines. Using the same weights as above, we get a similar diagram, which is not Voronoi for the same reason.

Thus, we can move the sites to \( \mathcal{A} \) and simply modify the weights to preserve the original power diagram on all of \( \mathcal{A} \).

B. EVERYTHING IS VORONOI

In this section, we give a stronger version of the uniqueness theorem, which asserts that under the same conditions as Theorem 4.1, not only is the mechanism unique up to positive-affine transformations, but it can be expressed as a positive-affine transformation of a classical peer prediction mechanism with respect to the Brier score. In other words, the maximal belief model constraint satisfying the conditions of the Theorem must in fact be a Voronoi diagram.\(^{12}\) For comparison, see Figure 6 for an example of a power diagram that is not a Voronoi diagram.

The intuition for this result is as follows. Because all pairs of cells have a nonempty border, the sites must be in general position, meaning that they are affinely independent within the affine hull of the simplex.\(^{13}\) Leveraging a standard geometric construction, we can compute the center \( \hat{p} \) of an \((m - 1)\)-sphere such that the sites all lie on the surface of the sphere (Figure 7M). This implies that the sites are all equidistant (in Euclidean distance) to center \( \hat{p} \), so we merely translate the sites so that \( \hat{p} \) coincides with the cell boundary intersection point \( p \) (Figure 7R). With appropriate weights, we know that this translation preserves the power diagram, yet now all sites are equidistant to \( p \) in Euclidean distance, so without loss of generality we can set \( w(s) = 0 \) for all \( s \) and the diagram is Voronoi.

\(^{12}\)We can see from Appendix A, as is well-known, every power diagram on an affine subspace of lower dimension (in our case, \( \{ v : v \cdot 1 = 1 \} \)) can be expressed as a Voronoi diagram in the full space by moving the sites perpendicularly to the affine subspace. Here we are saying something stronger: these diagrams are Voronoi for sites within the affine subspace.

\(^{13}\)Geometrically, no hyperplane in \( \mathbb{R}^m \) can contain all \( m \) sites.
Theorem B.1. Let $M$, $D$, and $p$ satisfy the assumption of Theorem 4.1. Then there exist points $p_1, \ldots, p_m$ in the affine hull of $D_m$ such that $M'$ is a truthful mechanism for $D$ if and only if

$$M'(x_i, x_j) = \alpha p^r(x_j) - \frac{\alpha}{2} ||p^r||^2 + f(x_j), \quad (9)$$

for some $f: [m] \to \mathbb{R}$ and $\alpha > 0$. In particular, $D$ is Voronoi.

Proof. Note that from Theorem 4.1 we already know that $D$ is a power diagram with sites $v^s \in \mathbb{R}^m$ and weights $w(s) \in \mathbb{R}$ for $s \in [m]$. Following [Eberly 2008], we will repeatedly refer to the identity

$$||a + c||^2 - ||b + c||^2 = ||a||^2 - ||b||^2 + 2c \cdot (a - b). \quad (10)$$

Let $d = m - 1$, and as outlined in Section A, we restrict attention to $\mathbb{R}^d$ via isometry to the affine hull of $\Delta_m$. Thus, we can assume WLOG that $v^1, \ldots, v^m, y \in \mathbb{R}^d$. Let $A$ be the $d \times d$ matrix whose $i$th row is given by $v^i - v^m$.

Note that $y$ is the unique point satisfying $y \in cl(cell(v^s))$ for all $s$, meaning that for some power distance $c \in \mathbb{R}$ we have

$$\forall s \in [m], \quad ||v^s - y||^2 - w(s) = c. \quad (11)$$

Let $w(m) = 0$ without loss of generality. Subtracting the equation for $s = m$ we have for all $s \in [d]$

$$0 = ||v^s - y||^2 - ||v^m - y||^2 - w(s) = ||v^s||^2 - ||v^m||^2 - 2y \cdot (v^s - v^m) - w(s), \quad (12)$$

where we have used the identity in Eq. 10. Letting $u \in \mathbb{R}^d$ be the vector with $u(s) = \frac{1}{2} (||v^s||^2 - ||v^m||^2 - w(s))$, and recalling the definition of $A \in \mathbb{R}^{d \times d}$, we see that $y$ is the unique solution to $Ay = u$, meaning that $A$ is invertible.

Let $\hat{u}$ be defined as $u$ but assuming $w(s) = 0$ for all $s$, i.e., with $\hat{u}(s) = \frac{1}{2} (||v^s||^2 - ||v^m||^2)$. Let $\hat{y} = A^{-1} \hat{u}$, the unique point satisfying $||v^s - \hat{y}||^2 = ||v^m - \hat{y}||^2$ for all $s \in [d]$. (One can verify this by unfolding $A\hat{y} = \hat{u}$ and applying Eq. 13.)

Geometrically, it is now clear that sites $\hat{v}^s = v^s + \hat{y} - y$ and weights $\hat{w}(s) = 0$ represent $D$, simply because $y$ is equidistant from $v + \hat{y} - y$ for all $s$, and as we only translated the sites and changed the weights, we can only translate the hyperplanes separating cells (i.e. we have not rotated any cell boundaries). To verify this claim algebraically, we will show that for any $q \in \mathbb{R}^d$, and any cells $s, s'$, the difference in power distances to $q$ remains the same:

$$\hat{v}^s - q||^2 - ||\hat{v}^{s'} - q||^2$$

$$= ||v^s + y - \hat{y} - q||^2 - ||v^{s'} + y - \hat{y} - q||^2$$

$$= ||v^s - \hat{y}||^2 - ||v^{s'} - \hat{y}||^2 + 2(y - q) \cdot (v^s - v^{s'})$$

$$= 2y \cdot (v^s - v^{s'}) - 2q \cdot (v^s - v^{s'})$$

$$= ||v^s - q||^2 - ||v^s - y||^2 - ||v^{s'} - q||^2 + ||v^{s'} - y||^2$$

$$= ||v^s - q||^2 - w(s) - ||v^{s'} - q||^2 + w(s')$$

where we have applied the identity of Eq. 10 three times, and in the final step we use Eq. 12. Thus, $D$ is a Voronoi diagram with sites $\hat{v}^s$. To complete the proof, we simply convert to a mechanism via Theorem 3.4 and then apply Theorem 4.1. □
Fig. 7: Converting a power diagram to a Voronoi diagram.